(12) 按照专利合作条约所公布的国际申请

(19) 世界知识产权组织
国 际 局

(43) 国际公布日
2019 年 7 月 4 日 (04.07.2019)

WIPO | PCT

(10) 国际公布号
WO 2019/127363 A1

(51) 国际专利分类号:
G06N 3/04 (2006.01)

(21) 国际申请号: PCT/CN2017/119821

(22) 国际申请日: 2017 年 12 月 29 日 (29.12.2017)

(25) 申请语言: 中文

(26) 公布语言: 中文

(71) 申请人: 清华大学(TSINGHUA UNIVERSITY) [CN/CN]; 中国北京市海淀区清华园 1 号, Beijing 100084 (CN)。

(72) 发明人: 张悠慧(ZHANG, Youhui); 中国北京市海淀区清华大学东主楼8区209室, Beijing 100084 (CN)。 季宇(JI, Yu); 中国北京市海淀区清华大学东主楼8区209室, Beijing 100084 (CN)。 张优扬(ZHANG, Youyang); 中国北京市海淀区清华大学东主楼8区209室, Beijing 100084 (CN)。

(74) 代理人: 北京睿邦知识产权代理事务所(普通合伙) (RUIBANG&PARTNERS); 中国北京市东城区夕照寺街4号东玖大厦B座409室, Beijing 100061 (CN)。

(81) 指定国(除另有指明，要求每一种可提供的国家保护): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL,

WO 2019/127363 A1

(54) **Title:** WEIGHT CODING METHOD FOR NEURAL NETWORK, COMPUTING APPARATUS, AND HARDWARE SYSTEM

(54) **发明名称**: 神经网络权重编码方法、计算装置及硬件系统

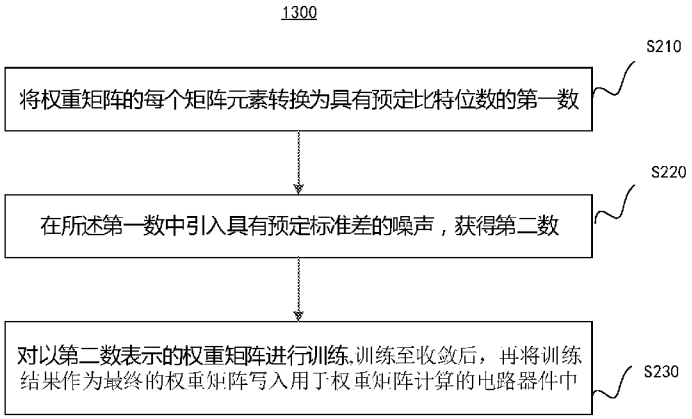


图 5

S210 Convert each matrix element of a weight matrix into a first number having a predetermined number of bits
S220 Introduce noise having a predetermined standard deviation into the first number to obtain a second number
S230 Train a weight matrix represented by the second numbers until convergence occurs, and then write the training result as a final weight matrix into a circuit device used for weight matrix computation

(57) **Abstract:** A non-assembling-based weight coding method for a neural network, comprising: a weight fixed-point conversion step of converting each matrix element of a weight matrix into a first number having a predetermined number of bits (S210); an error introduction step of introducing noise having a predetermined standard deviation into the first number to obtain a second number (S220); and a training step of training a weight matrix represented by the second numbers until convergence occurs, and then writing the training result as a final weight matrix into a single analog circuit device correspondingly representing one matrix element (S230), wherein a

[见续页]

SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW。

(84) **指定国**(除另有指明，要求每一种可提供的地区保护)：ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), 欧亚 (AM, AZ, BY, KG, KZ, RU, TJ, TM), 欧洲 (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG)。

本国际公布:

— 包括国际检索报告(条约第21条(3))。

single matrix element is represented by a single analog circuit device rather than multiple analog circuit devices assembled together. The coding method for a neural network can greatly reduce resource consumption without affecting the effect, thereby saving resource overhead, and thus, a large-scale neural network can be arranged with limited resources.

(57) 摘要：一种用于神经网络的非拼接权重编码方法，包括：权重定点化步骤，将权重矩阵的每个矩阵元素转换为具有预定比特位数的第一数（S210）；误差引入步骤，在所述第一数中引入具有预定标准差的噪声，获得第二数（S220）；和训练步骤，对以第二数表示的权重矩阵进行训练，训练至收敛后，再将训练结果作为最终的权重矩阵写入对应表示一个矩阵元素的单个模拟电路器件中（S230），其中，通过单个模拟电路器件而非多个模拟电路器件的拼接来表示单个矩阵元素。根据所述用于神经网络的编码方法，可以在不影响效果的情况下，极大的降低资源的消耗，从而节省资源开销，以在有限资源的条件下布置规模巨大的神经网络。

神经网络权重编码方法、计算装置及硬件系统

技术领域

本发明总体地涉及神经网络技术领域，更具体地涉及用于神经网络的权重编码方法、计算装置以及硬件系统。

背景技术

随着摩尔定律逐渐失效，现有芯片工艺进步放缓，人们不得不面向新应用和新器件。近年来，神经网络（Neural Network，NN）计算取得了突破性进展，在图像识别、语言识别、自然语言处理等诸多领域均取得了很高的准确率，但神经网络需要海量计算资源，现有的通用处理器已经很难满足深度学习的计算需求，设计专用芯片已经成为了一个重要的发展方向。与此同时，忆阻器的出现为神经网络芯片设计提供了一种高效的解决方案，忆阻器具有高密度、非易失、低功耗、存算合一、易于 3D 等优点，在神经网络计算中可以利用其阻值可调的特点作为可编程权重，并利用其存算合一的优点作高速乘加器。

神经网络组成单元均为神经元，由大量神经元相互连接成网络。神经元之间的连接可以看作带权重的有向边，神经元的输出会被神经元之间的连接所加权，然后传递给所连到的神经元，而每个神经元接收到的所有输入会被累加起来进行进一步处理，产生神经元的输出。神经网络的建模通常以若干神经元为一层，层与层之间相互连接来构建，图 1 所示的是一种链状的神经网络，图中每一个圆表示一个神经元，每一个箭头表示神经元之间的连接，每个连接均有权重，实际神经网络的结构不限于链状的网络结构。

神经网络的核心计算是矩阵向量乘操作，包含n个神经元的层L_n产生的输出可以用长度为n的向量V_n表示，与包含m个神经元的层L_m全相联，连接权重可以表示成矩阵$M_{n\times m}$，矩阵大小为n行m列，每个矩阵元素表示一个连接的权重。则加权之后输入到L_m的向量为$M_{n\times m}V_n$，这样的矩阵向量乘法运算是神经网络最核心的计算。

由于矩阵向量乘计算量非常大，在现有的通用处理器上进行大量的矩阵

乘运算需要耗费大量的时间，因此神经网络加速芯片也都是以加速矩阵乘法运算为主要的设计目标。忆阻器阵列恰好能胜任上述工作。首先 V 为一组输入电压，电压与忆阻器电导 G 相乘并叠加输出电流，输出电流与接地电阻 Rs 相乘得到输出电压 V'，整个过程在模拟电路下实现，具有速度快，面积小的优点。

然而使用基于忆阻器的芯片计算也存在精度低、扰动大，数模/模数转换开销大，矩阵规模受限等不足。而且虽然忆阻器可以高效地进行矩阵向量乘法运算，但是由于忆阻器芯片矩阵向量乘是在模拟电路中实现，所以不可避免的带来噪声和扰动，所以相对于神经网络，忆阻器的计算结果是不准确的。

由于忆阻器的工艺限制，使用忆阻器表示权重会有一定的误差。如图 3 所示，不同级的权重会有一定的重叠。为了避免重叠，现有方法一般是使用若干个低精度的忆阻器拼接来表示一个高精度权重，每个忆阻器精度都很低的情况下，可以认为权重数据是准确的。以用 2 个 2 比特忆阻器表示 4 比特权重为例，用一个 2 比特忆阻器来表示权重低 2 位，另一个表示高 2 位。

现有的ISAAC技术首先用浮点数训练一个神经网络，然后将权重数据“写入”忆阻器。ISAAC 是用 4 个 2 比特忆阻器器件来表示一个 8 比特的权重，这样可以利用更多的资源来提高矩阵运算精度。

ISAAC 使用拼接的方法来表示权重，效率比较低，需要很多的资源，比如表示一个 1 个权重，就需要 4 个忆阻器器件。

与ISAAC 类似，现有的PRIME 技术首先也用浮点数训练一个神经网络，然后使用 2 个 3 比特精度的输入电压来表示一个 6 比特输入，2 个 4 比特的忆阻器器件来表示一个 8 比特的权重，并且将正、负的权重分别用两组阵列表示。

PRIME 使用正负相加和高低位拼接方法来来表示权重，也需要大量的资源。即表示一个 1 个权重，就需要 4 个忆阻器器件。

基于忆阻器器件实现神经网络，必须克服权重读取误差的问题，这个问题是由器件特性和现有工艺造成的，目前难以避免。这些现有技术使用若干个低精度可以认为是“没有误差”的忆阻器拼接来表示一个高精度的权重，需要大量的资源，资源利用效率低。

因此，需要一种用于基于忆阻器神经网络的权重表示技术，以解决上述问题。

发明内容

鉴于上述情况，做出了本发明。

根据本发明的一个方面，提供了一种用于神经网络的非拼接权重训练方法，包括：权重定点化步骤，将权重矩阵的每个矩阵元素转换为具有预定比特位数的第一数；误差引入步骤，在所述第一数中引入具有预定标准差的噪声，获得第二数；和训练步骤，对以第二数表示的权重矩阵进行训练，训练至收敛后，得到训练结果，其中，所述训练结果将作为最终的权重矩阵，其各个矩阵元素被逐个写入对应表示一个矩阵元素的单个模拟电路器件中，其中通过单个模拟电路器件而非多个模拟电路器件的拼接来表示单个矩阵元素。

根据上述非拼接权重训练方法，在权重定点化步骤中，可以通过线性关系或对数关系进行第一数的转换。

根据上述非拼接权重训练方法，所述噪声可以为模拟电路的读写误差，并且服从正态分布规律。

根据上述非拼接权重训练方法，所述模拟电路器件可以为忆阻器、电容比较器或者电压比较器。

根据上述非拼接权重训练方法，所述第一数可以为定点数并且第二数可以为浮点数。

根据本发明的另一方面，提供了一种用于神经网络的非拼接权重编码方法，包括如下步骤：将权重矩阵的每个矩阵元素逐个写入对应表示一个矩阵元素的单个模拟电路器件中，以便通过单个模拟电路器件而非多个模拟电路器件的拼接来表示单个矩阵元素，其中，所述权重矩阵是通过上述非拼接权重训练方法得到的。

根据上述非拼接权重编码方法，在写入步骤之前，可以还包括如下步骤：权重定点化步骤，将权重矩阵的每个矩阵元素转换为具有预定比特位数的第一数；误差引入步骤，在所述第一数中引入具有预定标准差的噪声，获得第二数；和训练步骤，对以第二数表示的权重矩阵进行训练，训练至收敛后，得到训练结果。

根据本发明的另一方面，提供了一种神经网络芯片，具有通过模拟电路器件以硬件形式执行矩阵向量乘的操作的基本模块，其中，权重矩阵的每个矩阵元素被逐个写入对应表示一个矩阵元素的单个模拟电路器件中，以便通

过单个模拟电路器件而非多个模拟电路器件的拼接来表示权重矩阵的单个矩阵元素。

根据上述神经网络芯片，所述权重矩阵可以是上述非拼接权重训练方法得到的。

根据本发明的又一方面，提供了一种计算装置，包括存储器和处理器，存储器上存储有计算机可执行指令，所述计算机可执行指令当被处理器执行时执行根据上述非拼接权重训练方法或根据上述非拼接权重编码方法。

根据本发明的又一方面，提供了一种神经网络系统，包括：根据上述的计算装置；以及根据上述的神经网络芯片。

根据本发明，提供了一种用于神经网络的编码方法，可以在不影响效果的情况下，极大的降低资源的消耗，从而节省资源开销，以在有限资源的条件下布置规模巨大的神经网络。

附图说明

从下面结合附图对本发明实施例的详细描述中，本发明的这些和/或其它方面和优点将变得更加清楚并更容易理解，其中：

图 1 示出了链状的神经网络的示意图。

图 2 示出了基于忆阻器的交叉开关结构的示意图。

图 3 示出了在一个忆阻器上划分 8 级权重的权重统计分布图。

图 4 示出了根据本发明的神经网络的编码技术的应用情境的示意图。

图 5 示出了根据本发明的编码方法的总体流程图。

图 6 示出了使用现有高低位拼接方法和根据本发明的编码方法的实验效果对比。

具体实施方式

为了使本领域技术人员更好地理解本发明，下面结合附图和具体实施方式对本发明作进一步详细说明。

本申请提供一种新的编码方法（下文称为 RLevel 编码方法），其与现有方法的本质区别在于，新的编码方法并不要求使用单个器件表示的权重值不发生重叠，而是将这种误差引入训练当中。通过对含有噪声的权重矩阵进行训练，并使之能够训练至收敛，最后将收敛后的数值写入单个器件中，由此

既能够增强该模型的抗噪能力，也能够减少表示矩阵元素的器件的数量，降低了成本和资源消耗。

下面将结合附图对本申请的技术原理和实施方式进行详细分析。

图 3 示出了在一个忆阻器上划分 8 级权重的权重统计分布图。

如图 3 所示，由于忆阻器器件引起的误差近似于正态分布，假设器件的误差服从正态分布 $N(\mu, \sigma^2)$，如果用忆阻器电导值来表示一个 n 比特的值，则 μ 有2^n个可能的值。这里，本领域技术人员可以理解，为了简化计算，对应不同电导值 μ，采用相同的标准差 σ。

虽然，在下面的具体实施方式中以忆阻器作为示例进行说明，但是除了忆阻器以外的其他能够实现矩阵向量乘的电路器件也是可以的，比如电容或电压比较器。

根据正态分布叠加的性质：统计独立的常态随机变量 $X\sim(\mu_x,\sigma_x{}^2)$，$Y\sim(\mu_y,\sigma_y{}^2)$，那么它们的和也满足正态分布$U = X + Y\sim\left(\mu_x + \mu_y, \sigma_x{}^2 + \sigma_y{}^2\right)$。

假设如现有技术那样，用 2 个器件拼接来表示一个高精度权重。l 和 h 分别代表低位和高位器件，权重表示为$2^n * h + l$，低位和高位的误差分别是 $L\sim(l, \sigma^2)$，$H\sim(h, \sigma^2)$，则 $2^n * H\sim(h, 2^{2n} * \sigma^2)$。权重的取值范围是$2^{2n} - 1$，权重误差的标准差是$\sqrt{2^{2n} + 1}*\sigma$。我们把取值范围与标准差作为最终精度的标准，则拼接权重方法的精度为：$(2^{2n}-1) / (\sqrt{2^{2n} + 1}*\sigma)$。

相比，本申请中，用一个器件来表示高精度权重，其精度是 $(2^n-1) /\sigma$。

从上述结果可见，使用高低位拼接方法和单个器件表示权重的精度基本相同。

图 4 示出了根据本发明的神经网络的编码技术的应用情境的示意图。如图 4 所示，本公开的总体发明构思在于：将神经网络应用 1100 所采用的网络模型 1200 通过编码方法 1300 进行权重编码，将结果写入神经网络芯片 1400 的忆阻器器件，从而解决了基于忆阻器神经网络的权重表示需要大量器件的问题，最终在不明显损失精度的前提下，节省了大量的资源。

一、编码方法

图 5 示出了根据本发明的编码方法的总体流程图，包括如下步骤：

1、权重定点化步骤 S210，将权重矩阵的每个矩阵元素转换为具有预定比特位数的第一数；

根据硬件设计（单个忆阻器器件的精度需要硬件支持），在前向网络中，

将每一个权重值转换为一定精度的定点数，得到定点化权重。

这里，为了更好地说明本发明的方法，以下面的表 1 的 2*2 大小的权重矩阵 A 为例子进一步说明。

表 1 初始权重矩阵 A

0.2641	0.8509
0.3296	0.6740

当以 4 比特作为预定比特位数，将每个权重值转化为 4 比特定点数。其中，矩阵中的最大值 0.8509 对应于 4 比特的最大值，即 2^4-1=15，而其他值相应地进行线性换算而得到定点化权重，获得表 2 的定点数矩阵。

表 2 定点数矩阵 B

5.0000	15.0000
6.0000	12.0000

需要说明的是，上面是通过线性方式进行定点数的转换，但是本领域技术人员可以理解，也可以不通过线性方式，而通过对数或者其他计算方式进行转换。

2、误差引入步骤 S220，在所述第一数中引入具有预定标准差的噪声，获得第二数。

根据忆阻器器件特性，加入标准差为 σ 的正态分布的噪声进行训练，即权重 w=w+Noise，Noise~$(0, \sigma^2)$。需要说明的是，在这里，将第一数设定为定点数，而第二数等于第一数加噪音，因此第二数为浮点数。例如 0、1、2、3 四个定点数，加入了噪音，变成-0.1、1.02、2.03、2.88 四个浮点数，但是这样的设定并非限制性的，第一数也可以是浮点数。

3、训练步骤 S230，对以第二数表示的权重矩阵进行训练，训练至收敛后，再将训练结果作为最终的权重矩阵写入用于权重矩阵计算的电路器件中。

二、理论验证

下面给出实际的示例来从理论角度说明以同样的输入，使用根据本发明的RLevel编码方法的输出和根据现有技术的高低位拼接方法的输出有着接近的精度。

如果用两个 2 比特进行拼接，定点数矩阵 B（表 2）分解为高位矩阵 H

（表 3）和低位矩阵 L（表 4）：

表 3 高位矩阵 H

1.0000	3.0000
1.0000	3.0000

表 4 低位矩阵 L

1.0000	3.0000
2.0000	0.0000

拼接中，定点数矩阵 B 等于高位矩阵 H*4+低位矩阵 L，即 B=4*H+L，不管高位还是低位，最大值对应于 2 比特的最大值，即 3。

下面为了更好地模拟实际误差的引入，将定点数矩阵 B 分别按照 RLevel 方法和高低位拼接方法转换为 $4*10^{-6}$至 $4*10^{-5}$的电导值，则得到表 5 的 Rlevel 电导矩阵 RC、高位电导矩阵 HC 和低位电导矩阵 LC。

需要注意的是，根据本发明的训练过程并不会将矩阵转换为电导值，而是在第一数的基础上增加标准差为 σ 的正态分布的噪声进行训练。此处是为了说明，实际误差的引入是由于忆阻器器件或者其他所使用的电路器件在读取和写入过程中由于噪声和扰动而引起的，因此下面基于作为模拟值的电导值进行数据分析。

表 5 电导矩阵

Rlevel 电导矩阵 RC		高位电导矩阵 HC		低位电导矩阵 LC	
1.6000E-05	4.0000E-05	1.6000E-05	4.0000E-05	1.6000E-05	4.0000E-05
1.8400E-05	3.2800E-05	1.6000E-05	4.0000E-05	2.8000E-05	4.0000E-06

假设输入电压为：

0.10 0.15

【无噪声】

如果没有噪声，则基于上述输入电压，Rlevel 电导矩阵 RC、高位电导矩阵 HC 和低位电导矩阵 LC 的输出分别为：

表 6 电导矩阵输出

Rlevel 输出 RC_out	高位输出 HC_out	低位输出 LC_out	拼接输出 HLC_out
4.36000000E-06	4.0000E-06	1.0000E-05	2.18000000E-05
8.92000000E-06	5.8000E-06	4.6000E-06	4.46000000E-05

上述拼接输出为按照高位输出*4+低位输出。

如果将表 6 的结果，即 Rlevel 输出 RC_out 和拼接输出 HL_out，转换为 8 比特定点数进行比较，则可以看出两者均为：

125. 255.

【加入噪声】

如果对电导矩阵加入均值为 0 且标准差为 $0.05*4*10^{-5}$（即大约为 5%）的噪声，则得到表 7 的噪声矩阵。

表 7 噪声矩阵

Rlevel 噪声矩阵 RN		高位噪声矩阵 HN		低位噪声矩阵 LN	
1.6651E-05	4.4154E-05	1.9896E-05	3.9046E-05	1.6826E-05	3.9008E-05
1.9266E-05	3.0618E-05	1.5121E-05	4.0439E-05	3.0944E-05	1.4487E-06

仍然假设输入电压为：

0.10 0.15

则 Rlevel、高位和低位噪声矩阵输出分别为：

表 8 噪声矩阵输出

Rlevel 输出 RN_out	高位输出 HN_out	低位输出 LN_out	拼接输出 HLN_out
4.5550E-06	4.2578E-06	6.3242E-06	2.3355E-05
9.0081E-06	9.9704E-06	4.1181E-06	4.4000E-05

如果将表 8 的结果，即 Rlevel 输出 RN_out 和拼接输出 HLN_out，转换为 8 比特定点数进行比较，则可以看出两者分别为：

Rlevel 输出： 129.00 255.00

拼接输出： 135.00 255.00

由最终结果可见，无论加入噪声还是不加入噪声，根据本发明的 RLevel 编码方法都与现有技术的高低位拼接方法的输出有着非常接近的精度，因此，从理论角度验证了本发明的方案的实用性和可行性。

三、数据验证

为了从实验数据角度验证本发明的编码方法的有效性，申请人做了一系列的实验。

图 6 示出了使用现有高低位拼接方法和根据本发明的 RLevel 编码方法的实验效果对比。

本次实验使用卷积神经网络对 CIFAR10 数据集进行分类。该数据集有60000 个 32*32 像素的彩色图片，每张图片都属于 10 种分类之一。如图 6 所示，横坐标为权重精度，纵坐标为正确率。图中有两根线，下面的线是使用 RLevel 方法，分别用一个器件表示 2、4、6、8 比特的权重，而上面的线是以分别为 1、2、3、4 比特的 2 个器件拼接来表示 2、4、6、8 比特。

如图 6 所示，在本次实验中，用 RLevel 方法的准确率与用高低位拼接方法的准确率非常接近，但是由于只使用了一个器件，而不需要多个器件的拼接，属于非拼接编码，因此可以节省 50%的资源。如此，根据本发明的权重编码方法，不需要采用高低位拼接，就能够提供与现有的高低位拼接基本相同的精度，既解决了通过忆阻器等模拟电路来进行神经网络的权重矩阵计算需要布置大量电路器件的问题，也降低了成本，节省了资源。

以上已经描述了本发明的各实施例，上述说明是示例性的，并非穷尽性的，并且也不限于所披露的各实施例。在不偏离所说明的各实施例的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。因此，本发明的保护范围应该以权利要求的保护范围为准。

1、一种用于神经网络的非拼接权重训练方法，包括：

权重定点化步骤，将权重矩阵的每个矩阵元素转换为具有预定比特位数的第一数；

误差引入步骤，在所述第一数中引入具有预定标准差的噪声，获得第二数；和

训练步骤，对以第二数表示的权重矩阵进行训练，训练至收敛后，得到训练结果，

其中，所述训练结果将作为最终的权重矩阵，其各个矩阵元素被逐个写入对应表示一个矩阵元素的单个模拟电路器件中，其中通过单个模拟电路器件而非多个模拟电路器件的拼接来表示单个矩阵元素。

2、根据权利要求 1 所述的非拼接权重训练方法，其中，在权重定点化步骤中，通过线性关系或对数关系进行第一数的转换。

3、根据权利要求 1 所述的非拼接权重训练方法，其中，所述噪声为模拟电路的读写误差，并且服从正态分布规律。

4、根据权利要求 1 所述的非拼接权重训练方法，其中，所述模拟电路器件为忆阻器、电容比较器或者电压比较器。

5、根据权利要求 1 所述的非拼接权重训练方法，其中，所述第一数为定点数并且第二数为浮点数。

6、一种用于神经网络的非拼接权重编码方法，包括如下步骤：将权重矩阵的每个矩阵元素逐个写入对应表示一个矩阵元素的单个模拟电路器件中，以便通过单个模拟电路器件而非多个模拟电路器件的拼接来表示单个矩阵元素，

其中，所述权重矩阵是通过权利要求 1 到 5 中任一项所述的非拼接权重训练方法得到的。

7、根据权利要求 6 所述的非拼接权重编码方法，其中，在写入步骤之前，还包括如下步骤：

权重定点化步骤，将权重矩阵的每个矩阵元素转换为具有预定比特位数的第一数；

误差引入步骤，在所述第一数中引入具有预定标准差的噪声，获得第二

数；和

训练步骤，对以第二数表示的权重矩阵进行训练，训练至收敛后，得到训练结果。

8、一种神经网络芯片，具有通过模拟电路器件以硬件形式执行矩阵向量乘的操作的基本模块，

其中，权重矩阵的每个矩阵元素被逐个写入对应表示一个矩阵元素的单个模拟电路器件中，以便通过单个模拟电路器件而非多个模拟电路器件的拼接来表示权重矩阵的单个矩阵元素。

9、根据权利要求 8 的神经网络芯片，其中所述权重矩阵是通过权利要求 1 到 5 中任一项所述的非拼接权重训练方法得到的。

10、一种计算装置，包括存储器和处理器，存储器上存储有计算机可执行指令，所述计算机可执行指令当被处理器执行时执行权利要求 1 到 5 中任一项所述的非拼接权重训练方法或权利要求 6 到 7 中任一项所述的非拼接权重编码方法。

11、一种神经网络系统，包括：

如权利要求 10 所述的计算装置；以及

如权利要求 8-9 中任一项所述的神经网络芯片。

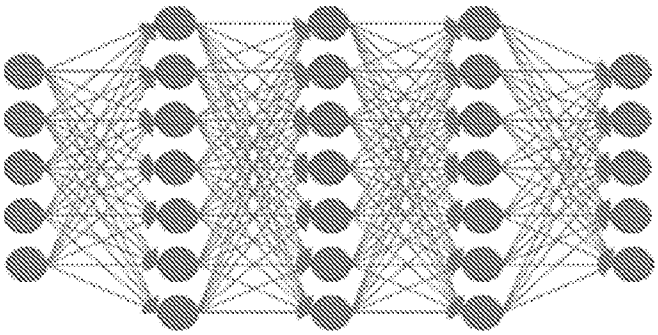


图 1

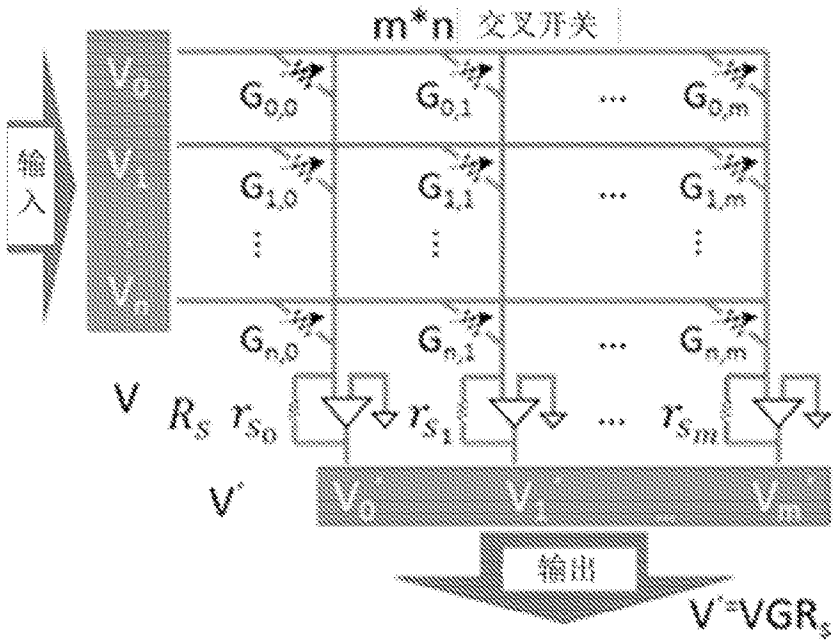

m*n 交叉开关
输入
V_0
V_1
V_n
$G_{0,0}$
$G_{0,1}$
$G_{0,m}$
$G_{1,0}$
$G_{1,1}$
$G_{1,m}$
$G_{n,0}$
$G_{n,1}$
$G_{n,m}$
V
R_S
r_{S_0}
r_{S_1}
r_{S_m}
V'
V_0'
V_1'
V_m'
输出
$V'=VGR_S$

图 2

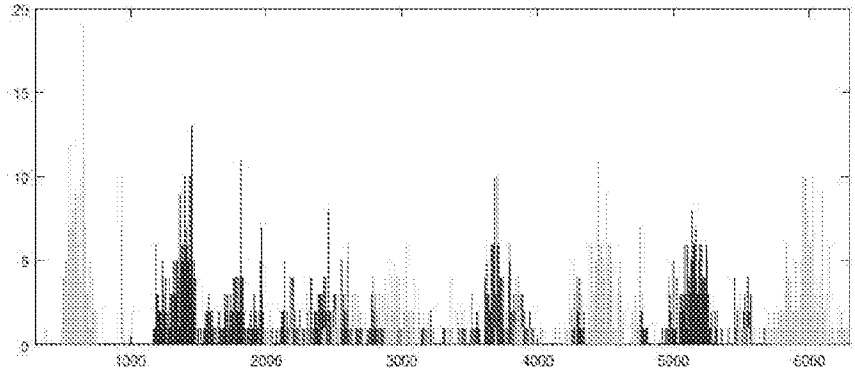

20
15
10
5
0
1000
2000
3000
4000
5000
6000

图 3

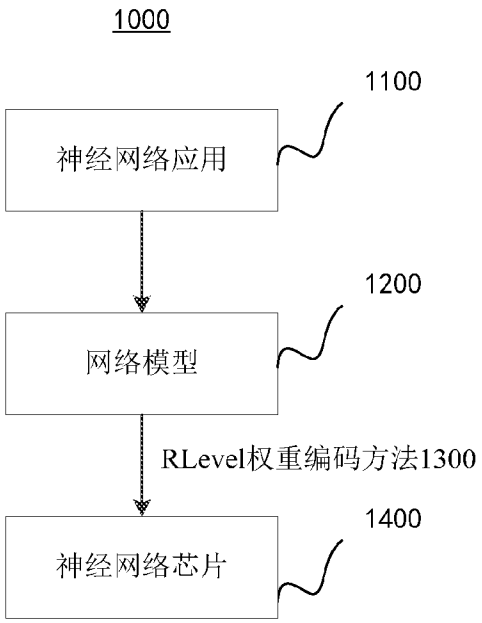

1000
1100
神经网络应用
1200
网络模型
RLevel权重编码方法1300
1400
神经网络芯片

图 4

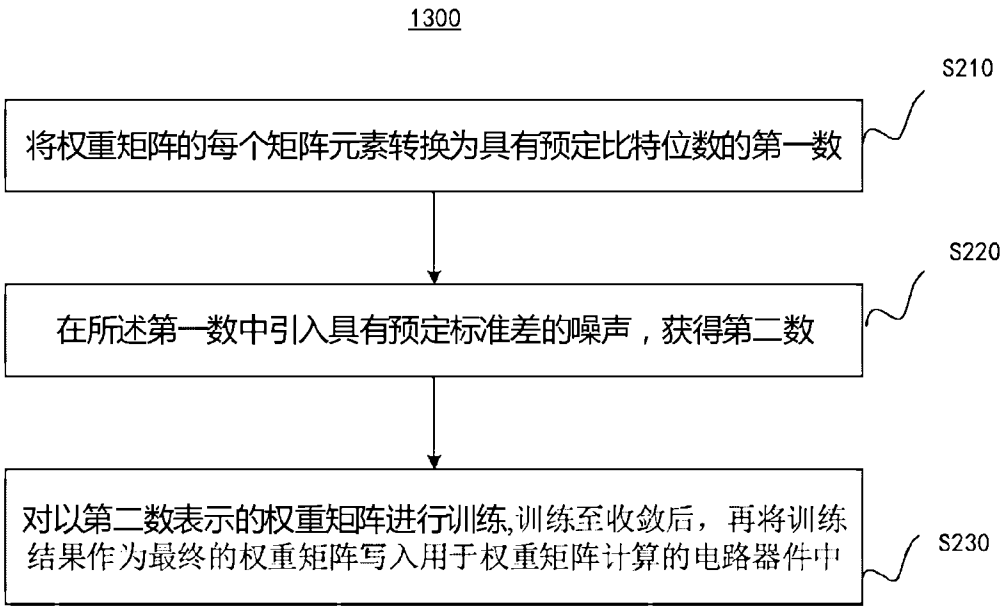


图 5

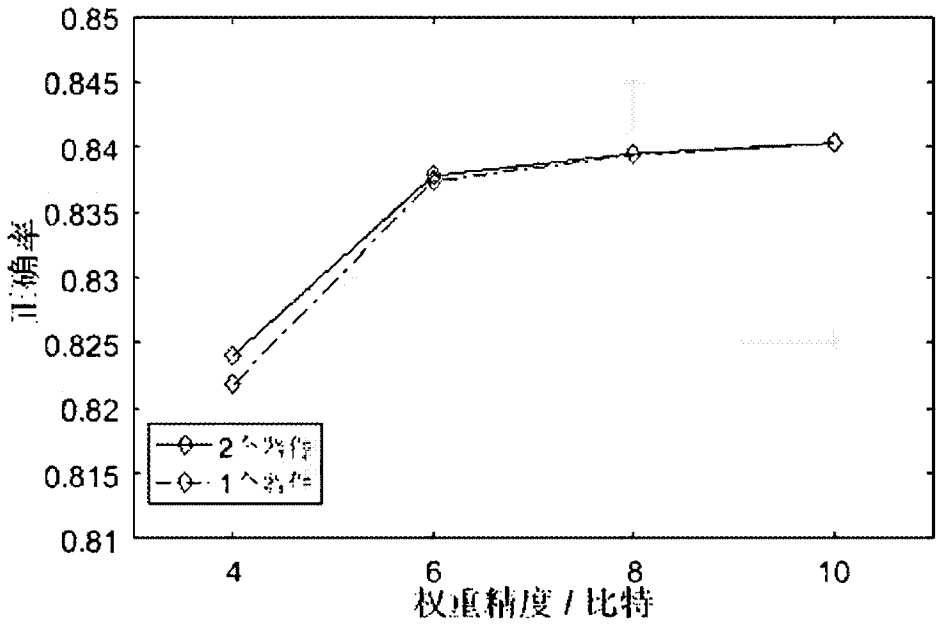


图 6

替换页（细则第26条）

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2017/119821

A. CLASSIFICATION OF SUBJECT MATTER

G06N 3/04(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNPAT, WPI, EPODOC, CNKI, IEEE: 神经网络, 忆阻器, 芯片, 重叠, 权重, 编码, 训练, 矩阵, 误差, 噪声, 定点数转换, 拼接, 模拟电路, neural w network, memristor, chip, overlap, weight, coding, train+, matrice, matrix, error, memristor+, noise, fixed w point, umber w conversion, splic+, analog w circuit

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 105224986 A (TSINGHUA UNIVERSITY) 06 January 2016 (2016-01-06) description, paragraphs [0005]-[0006] and [0019]-[0020]	1-11
A	CN 106650922 A (TSINGHUA UNIVERSITY) 10 May 2017 (2017-05-10) entire document	1-11
A	CN 107085628 A (SOUTHEAST UNIVERSITY) 22 August 2017 (2017-08-22) entire document	1-11
A	US 2017061281 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 02 March 2017 (2017-03-02) entire document	1-11

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
03 September 2018	**26 September 2018**

Name and mailing address of the ISA/CN	Authorized officer
State Intellectual Property Office of the P. R. China **No. 6, Xitucheng Road, Jimenqiao Haidian District, Beijing** **100088** **China**	
Facsimile No. **(86-10)62019451**	Telephone No.

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2017/119821

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN	105224986	A	06 January 2016	None	
CN	106650922	A	10 May 2017	None	
CN	107085628	A	22 August 2017	None	
US	2017061281	A1	02 March 2017	None	

国际检索报告

国际申请号
PCT/CN2017/119821

A. 主题的分类

G06N 3/04(2006.01)i

按照国际专利分类(IPC)或者同时按照国家分类和IPC两种分类

B. 检索领域

检索的最低限度文献(标明分类系统和分类号)

G06N

包含在检索领域中的除最低限度文献以外的检索文献

在国际检索时查阅的电子数据库(数据库的名称，和使用的检索词（如使用）)

CNPAT, WPI, EPODOC, CNKI, IEEE: 神经网络，忆阻器，芯片，重叠，权重，编码，训练，矩阵，误差，噪声，定点数转换，拼接，模拟电路，neural w network, memristor, chip, overlap, weight, coding, train+, matrice, matrix, error, memristor+, noise,fixed w point, umber w conversion, splic+, analog w circuit

C. 相关文件

类 型*	引用文件，必要时，指明相关段落	相关的权利要求
A	CN 105224986 A (清华大学) 2016年 1月 6日 (2016-01-06) 说明书第[0005]-[0006]、[0019]-[0020]段	1-11
A	CN 106650922 A (清华大学) 2017年 5月 10日 (2017-05-10) 全文	1-11
A	CN 107085628 A (东南大学) 2017年 8月 22日 (2017-08-22) 全文	1-11
A	US 2017061281 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 2017年 3月 2日 (2017-03-02) 全文	1-11

☐ 其余文件在C栏的续页中列出。

☑ 见同族专利附件。

* 引用文件的具体类型：

“A” 认为不特别相关的表示了现有技术一般状态的文件

“E” 在国际申请日的当天或之后公布的在先申请或专利

“L” 可能对优先权要求构成怀疑的文件，或为确定另一篇引用文件的公布日而引用的或者因其他特殊理由而引用的文件（如具体说明的）

“O” 涉及口头公开、使用、展览或其他方式公开的文件

“P” 公布日先于国际申请日但迟于所要求的优先权日的文件

“T” 在申请日或优先权日之后公布，与申请不相抵触，但为了理解发明之理论或原理的在后文件

“X” 特别相关的文件，单独考虑该文件，认定要求保护的发明不是新颖的或不具有创造性

“Y” 特别相关的文件，当该文件与另一篇或者多篇该类文件结合并且这种结合对于本领域技术人员为显而易见时，要求保护的发明不具有创造性

“&” 同族专利的文件

国际检索实际完成的日期	国际检索报告邮寄日期
2018年 9月 3日	2018年 9月 26日

ISA/CN的名称和邮寄地址	受权官员
中华人民共和国国家知识产权局(ISA/CN) 中国北京市海淀区蓟门桥西土城路6号 100088	李玉坤
传真号 (86-10)62019451	电话号码 53961358

国际检索报告
关于同族专利的信息

国际申请号
PCT/CN2017/119821

检索报告引用的专利文件			公布日（年/月/日）	同族专利	公布日（年/月/日）
CN	105224986	A	2016年 1月 6日	无	
CN	106650922	A	2017年 5月 10日	无	
CN	107085628	A	2017年 8月 22日	无	
US	2017061281	A1	2017年 3月 2日	无	